ECE 517: Machine Learning

Assignment 3.1: Theoretical Summary

David Kirby – 101652098 – davidkirby@unm.edu

Fall 2021

SCHOOL OF
ENGINEERING

This assignment is for the student to summarize the theoretical part of this module in a form that can be later inserted in an article. This assignment, then, will be used in next module as part of the corresponding assignment.

The only objective is to summarize the theory using a logic structure from the criterion to the implemented SVM. In next module, the assignment will be corrected to fit all the expectations of the theoretical part of a journal paper, and assessed following the corresponding rubric.

Outline of the assignment

Summarize the theory of this module in a maximum of three pages using the following structure:

1. Explain the concepts of Risk and Empirical Risk.
2. Explain the concepts of complexity and overfitting.
3. Introduce the concept of VC dimension.
4. Enunciate and interpret the VC theorem that describes the bound on the actual risk.
5. Introduce the SVM criteria.
6. Develop the analysis that leads to the dual solution of the SVM, and its main results.
7. Describe the properties of the Support Vectors.

1. Explain the concepts of Risk and Empirical Risk.

The concepts of risk and empirical risk are two methods to measure the accuracy and precision of a particular algorithm. In machine learning, there are generally three different sets of risks: actual, structural, and empirical. Using these we can determine how poor or well a given set of data can be represented by an algorithm. We are also able to determine if the algorithm will underfit or overfit our data. Finally, we can determine if the complexity of the model is too high or too low. Actual risk cannot be directly computed as we do not have the data a priori; instead, we try to approximate the actual risk with the help of the structural and empirical risks. Structural risk is designed to ensure that our algorithm does not become too specialized on a training dataset. It compares the complexity of the machine against its success of fitting the data, this in turn reduces the likelihood of overfitting or underfitting. The empirical risk; however, uses a given or known set of training data to analyze the prediction efficacy of the machine, for example through the minimization of the mean square. The empirical risk is the average loss over the data points.

2. Explain the concepts of complexity and overfitting.

In machine learning, complexity refers to the number of features and terms of a given model as well as its linearity. Complexity is directly proportional to the accuracy of the algorithm; however, increasing complexity can eventually reach a point of diminishing returns. When this happens, overfitting occurs and renders the model ineffective. Overfitting describes the condition when the algorithm classifies training data so closely that any test data could be misclassified. Over-complicating and overfitting create a highly specialized, but practically useless machine. We can reduce complexity by performing feature engineering and transforming data to extract valuable information.

3. Introduce the concept of VC dimension.

The Vapnik–Chervonenkis dimension was developed by Vladimir Vapnik and Alexey Chervonenkis during the 1960s–1990s as a statistical approach to classification problems. Used heavily in image classification, OCR, cancer prediction, and more, it is a binary classifier that attempts to maximize the separation between two classes of points. With regard to support-vector machines, the VC dimension determines the maximum number of vectors that can be shattered by a hyperplane and gives us a measure of the complexity of linear functions. If the VC dimension of an estimator is higher than the number of vectors to be classified, then the estimator is guaranteed to overfit if an empirical risk is minimized over the data, since all vectors will be correctly classified regardless of their statistical properties.

4. Enunciate and interpret the VC theorem that describes the bound on the actual risk.

With the Vapnik–Chervonenkis theorem, we need to define the linear empirical risk as:

$$R_{emp}(\boldsymbol{\alpha}) = \frac{1}{2N} \sum_{n=1}^{N} \left| \, y - f(\mathbf{x}, \boldsymbol{\alpha}) \, \right| \tag{1}$$

where $f(\cdot)$ is defined so that the loss function $\left| \, y - f(\mathbf{x}, \boldsymbol{\alpha}) \, \right|$ can only take the values of 0 or 1. Then, with the probability of $1 - \eta$, the following bound holds:

$$R(\boldsymbol{\alpha}) \leq R_{emp}(\boldsymbol{\alpha}) + \sqrt{\frac{h(\log(2N/h) + 1) - \log(\eta/4)}{N}} \tag{2}$$

This is a bound on the risk with probability $1 - \eta$, so it is therefore neither guaranteed nor dependent on the probability distribution. While the left side is not computable, the right one can be easily computed provided the knowledge of h, where the second term of the right side is called the structural risk (R_s). The inductive *Principle of Structural Risk Minimization* consists then on choosing a machine whose dimension h is sufficiently small, so that the bound on the risk is minimized.

5. Introduce the SVM criteria.

A support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. To construct the optimal hyperplane, it takes the support of two other hyperplanes that are parallel and equidistant from it on either side. These two support hyperplanes lie on the most extreme points between the classes and are called support-vectors. This optimal hyperplane is called the maximum margin hyperplane. The distance between the support hyperplanes is called the margin and the goal of SVM then is to find the maximum margin.

6. Develop the analysis that leads to the dual solution of the SVM and its main results.

The dual solution consists of minimizing the empirical risk and the structural risk through margin maximization as follows:

$$\text{minimize } L_p(\mathbf{w}, \xi_n) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n \tag{3}$$

$$\text{subject to } \begin{cases} y_n\left(\mathbf{w}^\top \mathbf{x}_n + b\right) > 1 - \xi_n \\ \xi_n \geq 0 \end{cases}$$

where C is a free parameter, p denotes primal, and ξ_n is the distance a datapoint exceeds the respective support hyperplane towards the other class, hence $\xi_n \geq 0$. The dual solution is designed to keep the objective function constant regardless of $\mathbf{w}$ and b. Optimization is performed by computing the gradient with respect to $\mathbf{w}$, and then nullified. The minima is directly found by solving derivatives of the objective function; however, since there are constraints, we need to first take the Langrangian of the function to solve for the minima. Finally, solving for $\mathbf{w}$ results in Karush–Kuhn–Tucker conditions and creates a saddle point, i.e., a global maximum over the domain of the criteria and a global minimum over the multipliers.

7. Describe the properties of the Support Vectors.

In support-vector machines, datapoints close to the hyperplane and satisfying certain conditions are called support vectors. If a datapoint is well within the boundary (support hyperplane), the penalizing factor ξ_n is 0. Otherwise, if the datapoint is on the other side, this factor ξ_n is equal to its distance between the datapoint and the support hyperplane, i.e., $\xi_n \geq 0$ and $\alpha_n = C$. If a sample is on the margin, $0 < \alpha_n = C$. Finally, if a sample is outside the margin, $\xi_n = 0$ and $\alpha_n = 0$.